R E P O R T R E S U M E S

ED 011 344 24 ED 011 344

TREATING DIVERSE MEASURES OF ABILITY IN INSTITUTIONAL RESEARCH.

BY- RALEY, WILLIAM L.
CALIFORNIA UNIV., BERKELEY, CTR.FOR R AND D IN ED

REPORT NUMBER BR-5-0248-3 PUB DATE OCT 66
CONTRACT OEC-6-10-106
EDRS PRICE MF-$0.09 HC-$0.28 7P.

DESCRIPTORS- *ABILITY, *ABILITY GROUPING, *RATING SCALES, *STATISTICAL ANALYSIS, *TEST RESULTS, *EQUATED SCORES, COMPARATIVE STATISTICS, TESTS OF SIGNIFICANCE, TEST INTERPRETATION, STATISTICAL DATA, RESEARCH AND DEVELOPMENT CENTERS, BERKELEY

A DISCUSSION WAS PRESENTED ON THE PROCEDURES OF EQUATING DIVERSE ABILITY SCORES OBTAINED BY DIFFERENT SCALED MEASURES. THE AUTHOR INDICATED THE NECESSITY OF EQUATING SUCH SCORES WHEN ABILITY IS TO BE TREATED AS A SINGLE INDEPENDENT VARIABLE IN SIGNIFICANCE TESTS AND IN COMPARATIVE ANALYSES OF DICHOTOMIZED GROUPS. A NEW RATIONALE FOR EQUATING ABILITY SCORES WAS DESCRIBED. IT WAS DESIGNED FOR A FORTHCOMING NATIONAL STUDY OF COMMUNITY COLLEGES TO MEET A SITUATION WHERE 10 DIFFERENT ABILITY TESTS WERE REPORTED BY THE PARTICIPATING INSTITUTIONS. THE RATIONALE SHOWED THAT ALL RAW ABILITY SCORES WOULD BE TRANSFORMED INTO PERCENTILES, OBTAINED FROM PUBLISHED NATIONAL NORMS FOR 13TH-GRADE COMBINED SEXES. AFTER PERCENTILES WERE OBTAINED, A CHART WAS PREPARED WHICH PERMITTED ASSIGNMENT OF ANY PERCENTILE TO AN APPROPRIATE STANINE, A SEGMENT OF A SCALE OF NINE. THESE STANINES WERE THAN CODED AS HIGH ABILITY, MIDDLE ABILITY, AND LOW ABILITY. THE HIGH-LOW GROUPING THAT RESULTED INCLUDED 23 PERCENT AT EACH END OF THE DISTRIBUTION. (JH)

Treating Diverse Measures of Ability in Institutional Research

William L. Raley

Center for Research and Development in Higher Education

University of California
Berkeley, California

October, 1966

A problem frequently encountered in research which uses samples obtained from a number of institutions is engendered by the fact that samples may have been assessed by different scaled measures of student "ability." When this problem exists, it is usually necessary to equate scores so that "ability" can be treated as a single independent variable. This can be done in one of the three following ways, depending on what questions are to be asked of the data.

I. If significance tests are to be performed:

Scores of the diverse instruments are converted to a common scale (e.g., ACE and AQT onto SCAT scales). The assumptions that are either explicit or implicit in this kind of operation are those concerned with parallel or alternate forms of tests. The recognized criterion for meeting the parallel forms assumption is, among other things, a correlation of .90 or above. The "other things" take the form of construct validity and the demonstration either statistically or logically that the tests are measuring the same factor or intellectual dimension. These assumptions can seldom be met with any degree of satisfaction when more than a few quite similar tests are dealt with and then only after intensive investigations involving correlational or analysis of variance techniques.

II. If groups are to be dichotomized for comparative purposes:

Scores are transformed to a normal distribution -- a less rigorous but more practical method than that described above. The main assumption in this approach is that tests administered by colleges (or by high schools for college admissions purposes) are, in a very broad

sense, measuring a trait that falls under the general rubric of "ability." While the intercorrelations of all tests involved are seldom known, it is assumed that the magnitude of the coefficients would be beyond chance. Another assumption is that the differences between norm groups as measured by the different tests are not so great as to invalidate the "ability" groupings to be formed for comparative purposes.

III. If groups are to be described:

Each test and its distribution is treated separately. No attempt is made to equate the scores from one test with any other. This avoids both compromise and criticism of the measurement theory or statistical techniques used. Unfortunately, however, it also introduces awkward analytical problems and tends to attenuate the results of a study.

In either of the first two approaches the investigator has to realize that a number of sources contribute to considerable error:

1. Differences in the degree to which the various tests measure "ability."
2. Sex differences.
3. Group differences.
4. Reliability of individual tests (e.g., error of measurement).

There is neither a body of literature nor a section of test theory exclusively devoted to the problem of equating scores from a number of instruments that measure a single trait. But research centers frequently encounter situations in which several tests, such as ability measures, must be handled as a single variable. The following rationale for transforming scores was devised for a forthcoming national study of community colleges, in order to meet a situation in which ten different ability

tests were reported by participating institutions. The situation was complicated by the fact that, while a number of institutions did use common tests, some reported only percentiles based on national norms, others reported local norms, still others reported national norms, but broken down by sex, and still others reported raw scores. A further complexity was added when institutions which used a common test reported scores based on different grade norms.

It was decided that under these circumstances ability as an independent variable within the study could be used only in a fairly gross fashion, and that all scores would be transformed into percentiles. These percentiles were obtained from published national norms for 13th grade combined sexes. After percentiles were obtained, a chart was prepared which permitted assignment of any percentile to its appropriate stanine. The top three stanines -- that is, 7, 8, and 9, which account for the top 23 per cent of the distribution -- were then coded as 1 = high ability. Stanines 4, 5, and 6 were combined and coded as 2 = middle ability, which accounts for 54 per cent in the middle part of the curve. Stanines 1, 2, and 3, accounting for the lower 23 per cent of the distribution, were coded 3 = low ability.

While there are numerous theoretical and statistical weaknesses in this approach, it offers two advantages which are difficult to discount.

1. The grossness of the stanines in groups of "threes" takes into account a great deal of the error known to exist, and also provides an objective basis for dealing with the "ability" groups.

2. The high-low grouping which resulted from the use of the stanines,

encompassing 23 per cent at each end of the distribution, closely follows Kelley's specification that the upper and lower 27 per cent of a normal distribution should be selected when forming dichotomous groups.